키워드 인식 시스템을 위한 연합 미세 조정 활용 위스퍼-타이니 모델

시바니 산제이 콜레카르, 김경백
전남대학교 인공지능학과
shivanikolekar@gmail.com, kyungbaekkim@jnu.ac.kr

Whisper-Tiny Model with Federated Fine Tuning for Keyword Recognition System

Shivani Sanjay Kolekar, Kyungbaek Kim
Dept. of Artificial Intelligence Convergence,
Chonnam National University.

Abstract

Fine-tuning is critical to enhance the model's ability to operate effectively in resource-constrained environments by incorporating domain-specific data, improving reliability, fairness, and accuracy. Large language models (LLMs) traditionally prefer centralized training due to the ease of managing vast computational resources and having direct access to large, aggregated datasets, which simplifies the optimization process. However, centralized training presents significant drawbacks, including significant delay, substantial communication costs, and slow convergence, particularly when deploying models to devices with limited resources. Our proposed framework addresses these challenges by employing a federated fine-tuning strategy with Whisper-tiny model for keyword recognition system (KWR). Federated learning allows edge devices to perform local updates without the need for constant data transmission to a central server. By selecting a cluster of clients and aggregating their updates each round based on federated averaging, this strategy accelerates convergence, reduces communication overhead, and achieves higher accuracy in comparatively less time, making it more suitable than centralized approach. By the tenth round of federated updates, the fine-tuned model demonstrates notable improvements, achieving over 95.48% test accuracy. We compare the FL-finetuning method with and centralized strategy. Our framework shows significant improvement in accuracy in fewer training rounds.

1. Introduction

The increasing reliance on voice-activated systems in real-world applications such as smart home devices, in-car assistants, and Internet of Things (IoT) environments demands efficient and accurate speech recognition models. Keyword spotting, a vital component of these systems, requires lightweight models that can operate effectively under strict computational constraints [5]. Large language models (LLMs), known for their remarkable capabilities in generating and processing text, have been increasingly applied to speech recognition tasks [1, 2]. However, the fine-tuning of these models for specific domains remains computationally expensive, especially in resource-constrained settings [2, 4].

Traditional centralized training methods aggregate vast amounts of data at a central server for model fine-tuning. While this approach is effective in high-resource settings, it is often impractical in distributed environments due to high latency, significant communication overhead, and privacy concerns associated with transmitting large volumes of data [4, 7]. Centralized training also struggles to adapt quickly to changing acoustic environments in edge devices such as smart speakers or in-car systems, where immediate and accurate keyword recognition is critical [5, 6].

To address these limitations, federated learning (FL) provides a decentralized alternative. FL allows edge devices to locally update model parameters based on their specific data, sending only the updated model weights back to a central server for aggregation [8]. This preserves privacy and reduces the need for high-bandwidth communication [9, 10]. In this work, we propose a federated fine-tuning approach for the Whisper-Tiny model, optimized for keyword recognition in resource-constrained environments. Using Federated Averaging (FedAvg), our method aggregates local model updates, improving model accuracy while minimizing communication and computational overhead [8].

Our FL fine-tuning framework efficiently reduces the time to convergence by selecting a subset of clients during each training round based on FedAvg strategy, which ensures faster adaptation client availability conditions. This work offers a scalable solution for deploying speech recognition models in

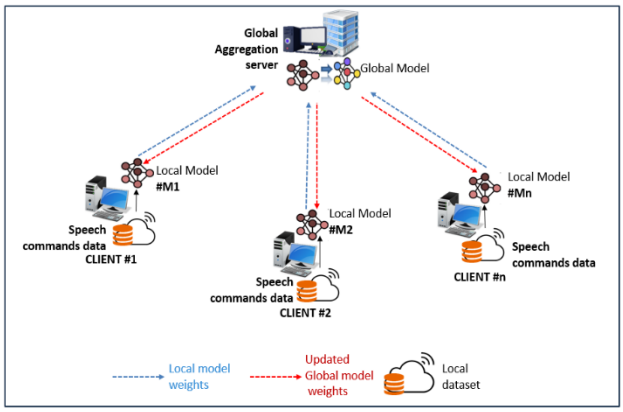

Figure 1: FL Training framework for fine-tuning of Whisper-tiny model for keyword recognition system

distributed environments where computational resources are limited [11].

2. Related Works

Large language models (LLMs), such as BERT and GPT, have demonstrated exceptional performance across a wide range of tasks, including natural language processing, speech recognition, and even multi-modal applications like visual question answering, owing to their large-scale pretraining on vast datasets [1,2]. These models, typically trained and fine-tuned in centralized environments, require significant computational resources and access to vast amounts of data, making them well-suited for high-performance computing settings [2,3]. However, in resource-constrained environments, such as smart home devices, IoT systems, or in-car assistants, centralized training introduces several challenges, including high latency, bandwidth costs, and privacy risks [4, 5].

Centralized training also limits the adaptability of LLMs in real-time keyword recognition tasks, which are crucial for enabling voice-activated functionalities on edge devices. These systems must handle diverse and dynamic acoustic environments, often dealing with background noise and varying voice profiles, which are difficult to manage with centralized models that do not continuously adapt to local conditions [6]. Moreover, transmitting large volumes of user data to central servers is often impractical in distributed systems, leading to increased communication costs and concerns about data privacy, particularly in sensitive applications such as healthcare and finance [4,7].

Federated learning (FL) has emerged as an efficient alternative to centralized training, offering a decentralized approach that enables edge devices to train models locally using their own data [8]. In FL, devices such as smartphones or IoT sensors update model parameters on-device and only transmit these model updates (i.e., weights) to a central server for aggregation, minimizing data transmission while preserving privacy [8,9]. This approach significantly reduces the communication burden and allows models to quickly adapt to local environments, making it particularly effective in environments with heterogeneous data and limited resources [10].

The Federated Averaging (FedAvg) algorithm, introduced by McMahan et al. [8], has become the standard in federated learning, efficiently aggregating client model updates to improve global model performance without the need to centralize raw data. However, despite these advances, fine-tuning large-scale models such as Whisper-Tiny for specific tasks, like keyword recognition, still presents significant computational challenges, even within the federated learning paradigm. This is especially true in resource-constrained environments where devices have limited computational power and storage capacity [8, 9].

In this paper, we propose a federated fine-tuning strategy for the Whisper-Tiny model, designed specifically to optimize keyword recognition in resource-constrained environments. Our approach builds on FedAvg, leveraging the concept of clustering clients to improve convergence rates and reduce computational overhead by selecting only a subset of clients to participate in each round of training [10]. By the tenth round of updates, our model achieves a test accuracy of 95.48%, significantly outperforming traditional centralized fine-tuning methods in terms of both efficiency and speed. This demonstrates that federated fine-tuning not only accelerates model convergence but also reduces communication costs and computational requirements, making it a scalable solution for deploying speech recognition models in distributed, resource-constrained environments.

3. Federated Fine-tuning of Whisper-tiny for Keyword Recognition Model (KWR)

In this section, we describe the Whisper-Tiny based Keyword Recognition (KWR) architecture, and the federated fine-tuning strategy implemented across five client devices to optimize the model for resource-constrained environments.

3.1 Whisper-Tiny Model for Keyword Spotting

The Whisper-Tiny model is a compact version of the Whisper architecture, a Transformer-based encoder-decoder model trained on 680k hours of labeled speech data through large-scale weak supervision [6]. Whisper-Tiny is designed for efficient inference with minimal computational resources, making it ideal for deployment in edge devices like smart home systems and in-car assistants.

Input Processing: The Whisper-Tiny model uses the Whisper Processor to pre-process audio inputs by converting them into log-Mel spectrograms. This feature extraction step is crucial for the model's ability to focus on important audio features during training and inference.

Transformer Architecture: The model utilizes self-attention mechanisms, which dynamically attend to different parts of the input audio sequence.

Embedding Space: The model's embedding layer maps input features into a higher-dimensional space optimized during training to facilitate the discrimination between different keywords.

3.2 Federated Fine-Tuning Strategy

To improve the performance of the Whisper-Tiny model in distributed environments, we employed a federated fine-tuning strategy using the Federated Averaging (FedAvg) algorithm. This strategy allows client devices (e.g., IoT devices, smart speakers) to fine-tune the model locally and share only model updates (weights) with a central server,

ensuring privacy preservation and minimizing data transmission.

Local Training on Clients: Five clients were selected to represent distributed devices, each holding a unique dataset with diverse audio conditions (e.g., different accents, noise levels). As shown in figure 1, at each federated round, each selected client trained the Whisper-Tiny model locally on its respective dataset for set number of epochs (10 epochs). The log-Mel spectrograms of the audio data were fed into the Whisper-Tiny model to update its embedding and attention layers for improved keyword recognition. After local training, the updated model weights were sent to the central server for aggregation.

Federated Averaging and Global Model Update

Once the clients completed local training, the server aggregated the model updates using the FedAvg algorithm. This process averaged the weights from the five clients, creating an updated global model that was then shared with all clients for the next training round for a total of 10 rounds.

This iterative process allowed the model to generalize well across diverse environments while retaining efficiency in communication and computation.

4. Evaluation

Dataset Description: The dataset for our keyword spotting system consists of audio recordings from the Google Speech Commands dataset, a widely used benchmark for training keyword models [12]. This dataset provides diverse acoustic conditions, ensuring robust model performance.

We used the latest version with over 30 different keywords. Of these, 10 primary commands—such as "Yes," "No," "Up," and "Down"—are utilized, while the remaining words help the model differentiate between commands and irrelevant speech. The dataset also includes noise recordings, improving the model's ability to handle background sounds in real-world environments.

Discussion: We used test accuracy and test loss as key evaluation metrics. Test accuracy measures the percentage of correct predictions made by the model on unseen data, offering a clear indication of how well the model generalizes beyond the training set.

In this paper, we evaluated the performance of the Whisper-Tiny model for keyword recognition using both centralized and federated (FedAvg) fine-tuning methods. As seen in Table 1, the federated fine-tuning approach significantly outperformed centralized fine-tuning, achieving a test accuracy of 95.48% compared to 92.94% for the centralized model (Figure2). Additionally, the federated method produced a lower test loss of 0.0040, in contrast to 0.0064 for the centralized approach. These results highlight the effectiveness of federated learning in improving model performance in resource-constrained environments.

5. Conclusion and Future Work

In this work, we proposed a federated fine-tuning strategy for the Whisper-Tiny model to enhance keyword recognition in resource-constrained environments. By leveraging federated strategy for fine-tuning a foundation model- Whisper-tiny, our approach efficiently aggregates local updates from distributed clients, reducing communication costs and computational overhead while accelerating convergence. The evaluation results demonstrated significant improvements in less time, achieving 95.48% test accuracy by the tenth round of updates, outperforming the centralized fine-tuning method. This framework provides a scalable, privacy-preserving, and resource-efficient solution for deploying speech recognition models in real-world applications. Future Work will explore optimizing communication efficiency further by integrating adaptive client selection and dynamic resource management. Additionally, applying this federated approach to other tasks, such as continuous speech recognition, can broaden the model's applicability in diverse edge environments.

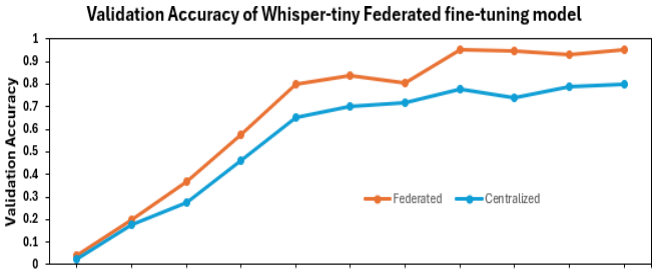

Figure 2: Fine-tuning of Whisper-tiny model validation accuracy for 10 rounds and a cluster of 5 clients.

Table 1: Evaluation: Test Accuracy and test Loss for Centralized and Federated (Fed-Avg) fine-tuning of whisper-tiny model for Keyword Recognition System

Evaluation of Federated fine-tuning of Whisper-tiny model for Keword Recognition System		
Fine-tuning method	**Test Accuracy**	**Test Loss**
Centralized	92.94	0.0064
Federated	95.48	0.0040

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government(MSIT). This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2024-00437718) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

[1]. Brown, T. B., et al. "Language models are few-shot learners." Advances in Neural Information Processing Systems 33 (2020): 1877-1901.

[2]. Devlin, J., et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[3]. Reddi, S. J., et al. "Adaptive Federated Optimization." International Conference on Learning Representations (2021).
[4]. Kairouz, P., et al. "Advances and open problems in federated learning." arXiv preprint arXiv:1912.04977 (2019).
[5]. Hinton, G., et al. "Deep neural networks for acoustic modeling in speech recognition." IEEE Signal Processing Magazine 29.6 (2012): 82-97.
[6]. McMahan, H. B., et al. "Communication-efficient learning of deep networks from decentralized data." Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. PMLR, 2017.
[7]. Bonawitz, K., et al. "Towards federated learning at scale: System design." Proceedings of the 2nd SysML Conference. 2019.
[8]. Li, X., et al. "On the Convergence of FedAvg on Non-IID Data." International Conference on Learning Representations (2020).
[9]. Konečný, J., et al. "Federated optimization: Distributed machine learning for on-device intelligence." arXiv preprint arXiv:1610.02527 (2016).
[10]. Smith, V., et al. "Federated multi-task learning." Advances in Neural Information Processing Systems 30 (2017): 4424-4434. Link
[11]. Wang, S., et al. "Adaptive Federated Learning in Resource-Constrained Edge Computing Systems." IEEE Journal on Selected Areas in Communications 37.6 (2019): 1205-1221.
[12]. Warden, Pete. "Speech commands: A dataset for limited vocabulary speech recognition." arXiv preprint arXiv:1804.03209 (2018).